Navigating and Analyzing Large Economic Datasets

Xiaoyi Zhao

Virginia Tech Data Science for the Public Good Program

xiaoyiz@vt.edu

VIRGINIA TECH

Table of Contents

VIRGINIA TECH

What is Census data

- Census data are official statistics collected systematically by governments to capture demographic, economic, and social information about the population.
- Typically collected at regular intervals
- Includes information on population size, age, gender, income, employment status, education, housing conditions, and more.

Why do researchers use census data?

- Wide Coverage and Representativeness
- Reliability and Credibility:
- Supports a wide range of information that supports research in different topics
- Enables researchers to study changes and trends over time.

Where to Find Census Data

- **Public Agencies**; for example:
  - **Bureau of Labor Statistics**: Employment, wages, unemployment rates (CPS, QCEW, LAUS)
  - **Census Bureau**: Population demographics, economic status, housing (ACS, Decennial Census, SIPP)
  - **Federal Reserve**: Financial data, consumer surveys (FRED, Survey of Consumer Finances)
- **International Organizations**
- **State government**: State-specific data such as employment, education, health, demographics
- **Academic Data Archives**: IPUMS
- **Direct Data Requests**

VIRGINIA TECH

How to identify information in the datasets

- **Geographical and temporal coverage**
  - National, regional, state, or local levels
  - Annual, quarterly, monthly, or one-time collection
- **Variable Definitions**

  Example: Age group categories, inflation-adjusted prices
- **Survey Design and Dataset Type**
  - Cross-sectional vs. longitudinal
  - Sample design (e.g., random, weighted)

Tips for choosing the correct dataset

- Clearly figure out the main variables needed for your topic.
- Decide what geographic area (national, state, county) and time period you're interested in.
- Look for unique identifiers (like FIPS codes) to easily merge datasets.
- Check if the dataset is easy to access, user-friendly, and available in the format you prefer.

Accessing the Data

- Option 1: Download Directly from Websites
- Option 2: Use R Packages and APIs.

Things to Check When Downloading Data from Websites

- Make sure it's compatible with your tools (CSV, Excel, Stata, etc.).
- Check if the data has been edited or allocated.
- Be aware of Missing value
- Be aware of Time Coverage and Consistency

Example 1: Extract Data from Ipums

- This will be a live walkthrough of how to search and extract data from IPUMS.
- **Goal of the Demo**:
  - Demonstrate how to collect data to study the relationship between years of education and income.
  - Walk through searching for relevant variables (e.g., education level, income).
  - Show how to select samples and submit a data extract.
- **Tip**: To use IPUMS, make sure you're logged in to your free IPUMS account before starting.

VIRGINIA TECH

Option 2: Use R Packages and APIs

- API (Application Programming Interface): In simple terms, an API is like a personal key that helps two programs connect and share information automatically
- Let's walk through an example of how to request an API key and use it to pull data from the BLS into R

Example 1: Deschênes and Greenstone (2007)

- Title: The Economic Impacts of Climate Change: Evidence from Agricultural Output and Random Fluctuations in Weather
- Research Question focus: Links between farmland productivity and climate.
- Dataset:
  1. U.S. Census of Agriculture (county-level agricultural data)
  2. National Resource Inventory (for soil quality data)
  3. Parameter-Elevation Regressions on Independent Slopes Model (PRISM) (for Climate and Weather Data)

VIRGINIA TECH

Issue 1:Choosing Between 1-Year vs. 5-Year ACS Estimates

- Challenges: The ACS provides both 1-year and 5-year estimates. Which one is more appropriate for your project?
-
  - **Use 1-year estimates** for timely analysis, especially when evaluating recent changes (e.g., after a policy shift).
  - **Use 5-year estimates** when you need more reliable and stable data, especially for small populations or geographies.

Source: U.S. Census Bureau – American Community Survey (ACS) 1-Year vs. 5-Year Estimates. mySidewalk Knowledge Center.

VIRGINIA TECH

Issue 2: Variables/Years May Not Match Exactly

- Challenges: Sometimes the variables/years available in the dataset don't match exactly what we want due to how the survey design.
-
  - Look at how similar papers or studies have addressed the issue.
  - For example, if you want "years of education" but only have a categorical variable for highest degree attained (e.g., grouped into intervals), some researchers recode it and assigning average years to each category.

(Henderson et al., 2011)

Thank You for listening!

Questions and Discussion.

Reference

Deschênes, O., Greenstone, M. (2007). The Economic Impacts of Climate Change: Evidence from Agricultural Output and Random Fluctuations in Weather. American Economic Review, 97(1), 354–385. https://doi.org/10.1257/aer.97.1.354

Henderson, D. J., Polachek, S. W., Wang, L. (2011). Heterogeneity in schooling rates of return. Economics of Education Review, 30(6), 1202–1214. https://doi.org/10.1016/j.econedurev.2011.05.002

U.S. Census Bureau American Community Survey (ACS) 1-Year vs. 5-Year Estimates | mySidewalk Knowledge Center. (n.d.). Retrieved June 3, 2025, from https://help.mysidewalk.com/en/articles/2581167-u-s-census-bureau-american-community-survey-acs-1-year-vs-5-year-estimates